Project: Bank Marketing (Campaign)

Week 8 deliverables

Group name: Fleifel-solo

Name: Rania Tarek Fleifel

Email: raniatarekfleifel@gmail.com

Country: Egypt

College: Cairo university Faculty of engineering

Specialization: Data science

Internship Batch: LISUM13: 30

Submission date: 23rd November 2022

Problem description

Provide ABC bank with a model that enables them to target costumers who're more probable to invest in their new term deposit product.

Business understanding

As ABC launches their new term deposit product, they need their outreach teams to effectively market the product to costumers whose interactions with the bank (loans, responsiveness to offers, etc) as well as their personal standing (job stability, marital status, age, etc) show high possibility of purchasing the deposit product. The need to focus on those costumers is -at the heart of it- the bank's strategy to effectively use the marketing team's resources to spread the deposit product among interested customers.

Deposit products promise costumers a high interest rate in return to locking an amount of their money for some time. Many factors decide whether a costumer invest in such product or not. The most important is his standing in life in general. For instance, customers who have savings beyond their day-to-day spending and have sitting-money would seemingly fit the profile of a perfect costumer. If a costumer's age is 60+, he's not a very good fit since he has limited resources and is retired, hence has no income except his pension which might not offer any excess to invest in this product. Customers who are used to taking loans could benefit as well if the interest rate from their savings covers their installments to the bank. Jobs play an important role as well in defining whether a costumer is a good fit. Doctors, engineers and similar prestigious jobs that are known to pay well are good candidates, as well as individuals with a long-standing job; 20+ work in a certain company shows stability.

What type of data you have got for analysis?

A- Will I use the older version datasets or the newer version datasets?

There are two versions of the data available. The newer version has more features/columns (21 vs. 17). However, the older version has more unique data points (45211 vs. 41176). Ideally, I would analyze the features' importance and relation to the output "y" and choose the dataset that provide more descriptive values where it matters (aka the "influential" features). However, in accordance to the deliverables of this project, the decision should be made before starting the analysis.

In this subsection we look into what each version offers and conclude with the version we will proceed with. To make an informed decision, we first begin with contextual understanding of the variables as shown in the bellow table.

	Variable	Contextual remarks
Numeric	age	Age of costumer at most recent contact date
	balance	Average amount in euros
	*day (old dataset)	Last contact day of the month
	duration	Last contact duration in seconds
	campaign	Number of contacts with this client for this campaign
	previous	Number of contacts with this client before this campaign
	pdays	Number of days since the client was last contacted on a previous campaign *999: not previously contacted (new dataset)
	*emp.var.rate (new dataset)	Employment variation rate, with a quarterly frequency
	*cons.price.idx (new dataset)	Monthly average consumer price index
	*cons.conf.idx (new dataset)	Monthly average consumer confidence index
	*euribor3m (new dataset)	Daily three-month Euribor rate
	*nr.employed (new dataset)	Quarterly average of the total number of employed citizens
Categorical	job	Occupation of costumer
	marital	Marital status of costumer (married,divorced,single) p.s divorced=divorced/widowed
	education	Education level of costumer
	*Day_of_week (new dataset)	Last contact day of the week
	month	Last contact month
	default	If the client has credit in default?
	housing	If the client has a house loan contract (yes/no)
	loan	If the client has a personal loan contract (yes/no)
	contact	Last contact communication type
	poutcome	Outcome of previous campaign
	y	Did the client subscribe for client deposit?

Next, we look into the discrepancies in the features of the two datasets and their values, we look into this for numerical and categorical features separately:

1) The discrepancies in numerical features

int variable "emp.var.rate" in new version
~~categorical variable "day_of_week" in new version~~
int variable "cons.price.idx" in new version
int variable "nr.employed" in new version
int variable "cons.conf.idx" in new version
int variable "euribor3m" in new version
int variable "balance" in old version
~~int variable "day" in old version~~

The following remarks are drawn:

- The variables “day” and “day_of_week” serve the same purpose
- The 5 variables “emp.var.rate”, “cons.price.idx”, “cons.conf.idx”, “nr.employed”, “euribor3m” are related to economic indicators
- The variable “balance” shows the numeric average yearly balance. It could be an indicative of how lucrative a costumer’s money is and hence how probable he would invest.

2) The cardinality in categorical features

```
contact non-common values= {'unknown'}
contact 's unique: ['cellular', 'telephone']
contact 's Old unique: ['cellular', 'telephone', 'unknown']

default non-common values= {'unknown'}
default 's unique: ['no', 'unknown', 'yes']
default 's Old unique: ['no', 'yes']

y cardinaity match!

poutcome non-common values= {'other', 'unknown', 'nonexistent'}
poutcome 's unique: ['failure', 'nonexistent', 'success']
poutcome 's Old unique: ['failure', 'other', 'success', 'unknown']

loan non-common values= {'unknown'}
loan 's unique: ['no', 'unknown', 'yes']
loan 's Old unique: ['no', 'yes']

month non-common values= {'feb', 'jan'}
month 's unique: ['apr', 'aug', 'dec', 'jul', 'jun', 'mar', 'may', 'nov', 'oct', 'sep']
month 's Old unique: ['apr', 'aug', 'dec', 'feb', 'jan', 'jul', 'jun', 'mar', 'may', 'nov', 'oct', 'sep']

education non-common values= {'basic.9y', 'basic.4y', 'tertiary', 'illiterate', 'university.degree', 'primary', 'high.school', 'professional.course', 'basic.6y', 'secondary'}
education 's unique: ['basic.4y', 'basic.6y', 'basic.9y', 'high.school', 'illiterate', 'professional.course', 'university.degree', 'unknown']
education 's Old unique: ['primary', 'secondary', 'tertiary', 'unknown']

housing non-common values= {'unknown'}
housing 's unique: ['no', 'unknown', 'yes']
housing 's Old unique: ['no', 'yes']

marital non-common values= {'unknown'}
marital 's unique: ['divorced', 'married', 'single', 'unknown']
marital 's Old unique: ['divorced', 'married', 'single']

job cardinaity match!
```

The following remarks are drawn:

- The new dataset contains more elaborate info about the education of costumers
- The new dataset doesn’t take months “jan” and “feb” in consideration at all.
- The new dataset shows the value “unknown” in a number of features “housing, marital”, “loan” and “default”
- The old dataset shows ambiguous values such as “unknown” and “other” in the features “poutcome” and “contact”.

According to this short analysis, **we decide to proceed with the “Old dataset”**. This decision is in favor of the dataset that:

- Has more data points,
- Covers the whole period of analysis (all months),
- Has less ambiguous values that could later affect the predictive model
- Avoids high cardinality in some features (education).
- Has variables (day, month) which can be used to impute the 5 features () that represent economic indicators

B- Data univariate analysis:

Numeric variables:

Variable	Statistic/calculation

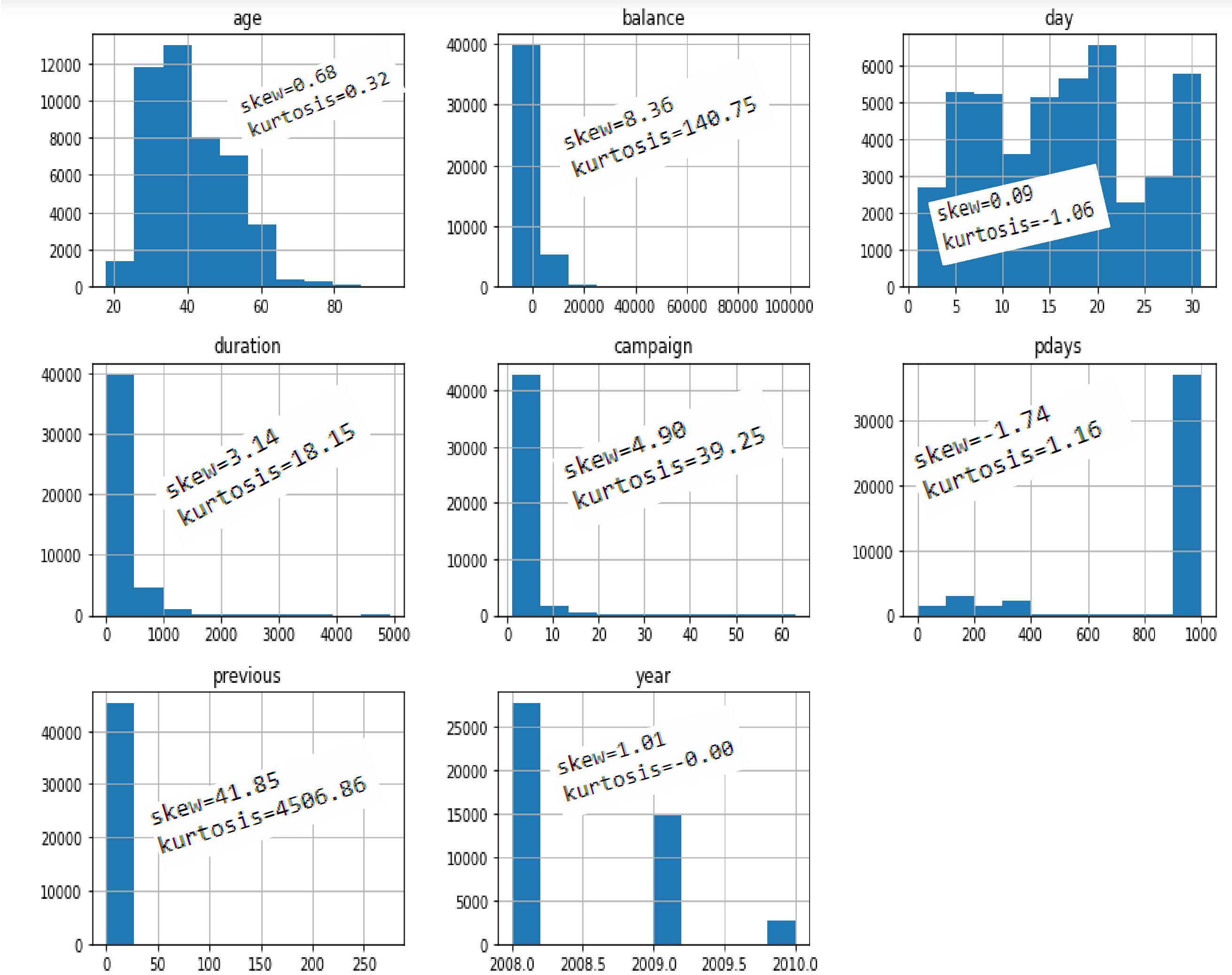
Age	`dtype=int64, #of nulls=0, #of zeros=0`
	`Q1=33.0, Q2=33.0, Q3=33.0, IQR=0.0`
	`min=18.0, max=95.0, range=range(18, 95)`
	`mean=40.94, std=10.62, median=39.00, mad=8.74`

Balance	`dtype=int64, #of nulls=0, #of zeros=3514`
	`Q1=72.0, Q2=72.0, Q3=72.0, IQR=0.0`
	`min=-8019.0, max=102127.0, range=range(-8019, 102127)`
	`mean=1362.27, std=3044.77, median=448.00, mad=1551.51`

*day (old dataset)	`dtype=int64, #of nulls=0, #of zeros=0`
	`Q1=8.0, Q2=8.0, Q3=8.0, IQR=0.0`
	`min=1.0, max=31.0, range=range(1, 31)`
	`mean=15.81, std=8.32, median=16.00, mad=7.06`

Duration	dtype=int64, #of nulls=0, #of zeros=3 Q1=103.0, Q2=103.0, Q3=103.0, IQR=0.0 min=0.0, max=4918.0, range=range(0, 4918) mean=258.16, std=257.53, median=180.00, mad=170.97
Campaign	dtype=int64, #of nulls=0, #of zeros=0 Q1=1.0, Q2=1.0, Q3=1.0, IQR=0.0 min=1.0, max=63.0, range=range(1, 63) mean=2.76, std=3.10, median=2.00, mad=1.79
Previous	dtype=int64, #of nulls=0, #of zeros=36954 Q1=0.0, Q2=0.0, Q3=0.0, IQR=0.0 min=0.0, max=275.0, range=range(0, 275) mean=0.58, std=2.30, median=0.00, mad=0.95
Pdays	dtype=int64, #of nulls=0, #of zeros=0 Q1=999.0, Q2=999.0, Q3=999.0, IQR=0.0 min=1.0, max=999.0, range=range(1, 999) mean=857.57, std=303.25, median=999.00, mad=231.21
**year	*To be deduced*
**emp.var.rate	*To be merged from the new dataset*
**cons.price.idx	
**cons.conf.idx	
**euribor3m	
**nr.employed	

Histogram of numeric attributes:



Issues with the numeric variables and how to resolve them:

- The ‘pdays’ value 999 that represent “not previously contacted for a previous campaign” is over-riding the actual statistics of the variable (outlier)

  ➔ *Represent this information with a different value for an existing variable or a new variable.*

  ➔ *Filter the attribute before 999*

- The ‘year’ should be added in order to impute the economic indicator correctly. We use the information that the data is ordered from 2008 to 2010 to generate this attribute

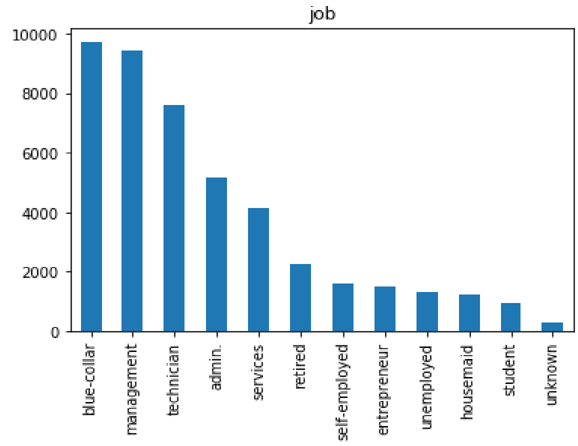
  ➔ *This requires manipulating the new dataset to have ‘day’ attribute to correctly merge euribor3m that is updated daily*

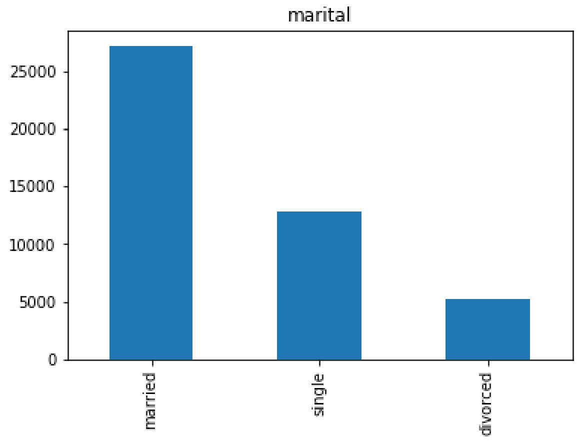
- The 5 economic indicators will be imputed from the new dataset (the timeseries indicators will have missing values at dates that don't exist in the new dataset)

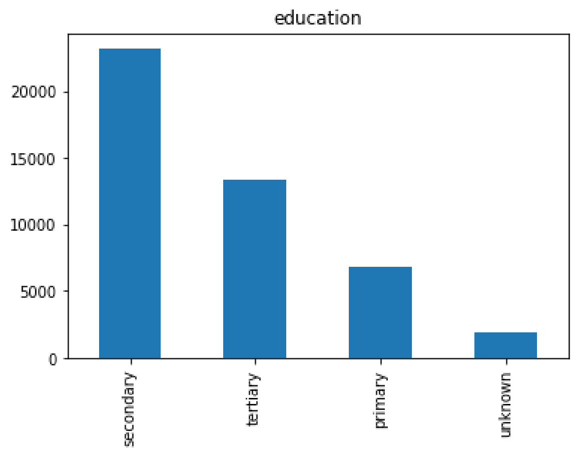
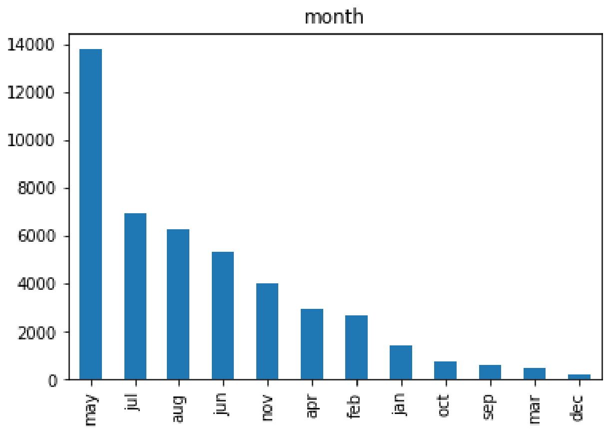
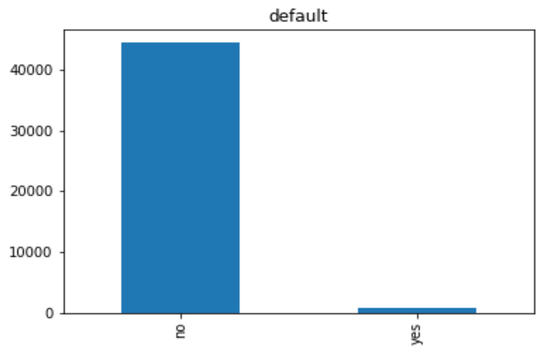
  ➔ *Conclude the missing data based on mean/mode/median or impute the data from a different source*

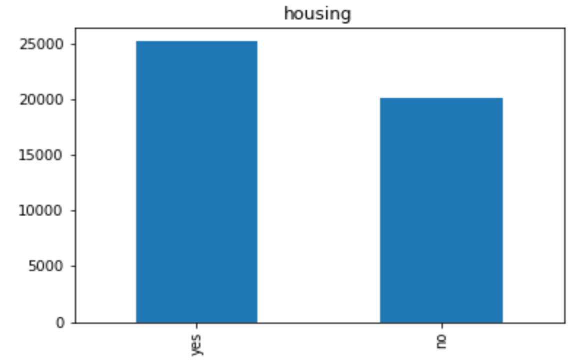
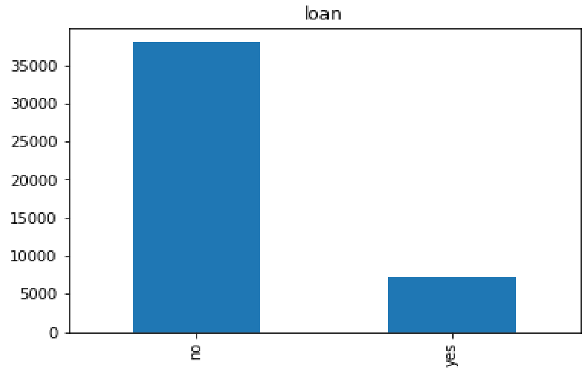
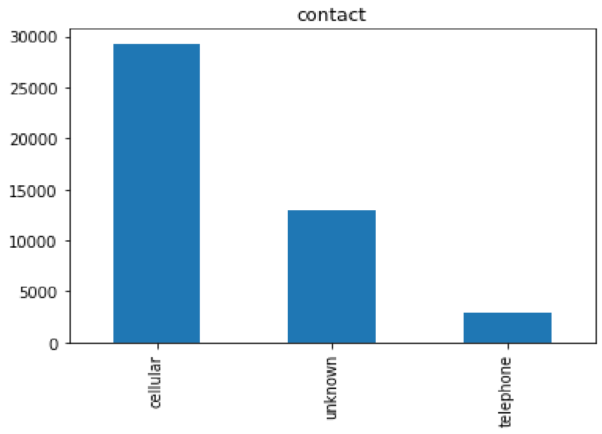
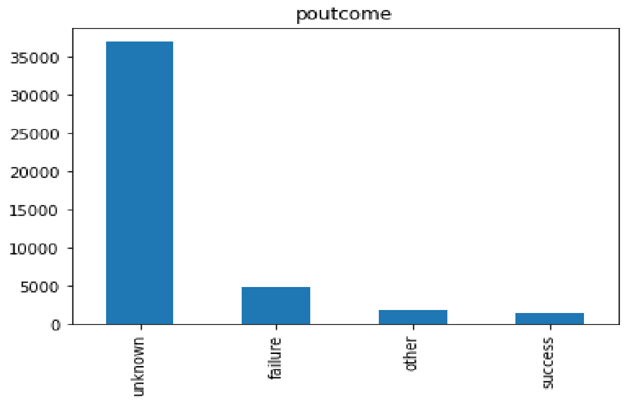
- **Categoric variables**

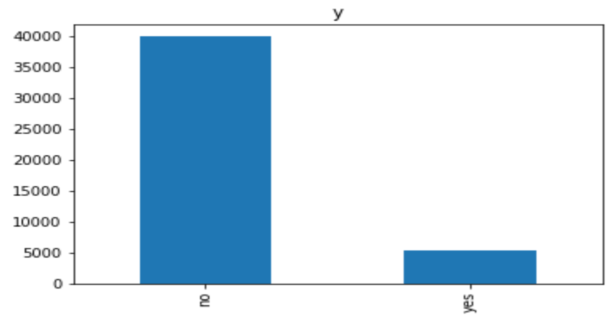
<table>
<tr>
<td>Job</td>
<td>dtype=object
#of nulls|zeros=0

management(9458)
blue-collar(9732)
retired(2264)
unknown(288)
technician(7597)
admin.(5171)
unemployed(1303)
services(4154)
student(938)
entrepreneur(1487)
self-employed(1579)
housemaid(1240)</td>
<td>
</td>
</tr>
<tr>
<td>marital</td>
<td>dtype=object
#of nulls|zeros=0

divorced(5207)
married(27214)
single(12790)</td>
<td>
</td>
</tr>
</table>

education	dtype=object #of nulls|zeros=0 secondary(23202) primary(6851) tertiary(13301) unknown(1857)	
month	dtype=object #of nulls|zeros=0 jun(5341) sep(579) aug(6247) jan(1403) oct(738) dec(214) feb(2649) jul(6895) apr(2932) nov(3970) mar(477) may(13766)	
default	dtype=object #of nulls|zeros=0 yes(815) no(44396)	

housing	dtype=object #of nulls|zeros=0 yes(25130) no(20081)	
loan	dtype=object #of nulls|zeros=0 yes(7244) no(37967)	
contact	dtype=object #of nulls|zeros=0 cellular(29285) telephone(2906) unknown(13020)	
poutcome	dtype=object #of nulls|zeros=0 success(1511) failure(4901) other(1840) unknown(36959)	

y	`dtype=object`
	`#of nulls|zeros=0`
	
	`yes(5289)`
	`no(39922)`
		

<u>Issues with the categoric variables and how to resolve them:</u>

- Imbalance of categorical target (class imbalance). This causes the predictive model to favor ‘yes’ regardless of the predictive model because it’s the majority class.

  ➔ *Undersample the majority class/oversample the minority class*

  ➔ *Use precision and recall for evaluating the predictive model and not accuracy*

- The high cardinality of jobs might cause non-conclusive results

  ➔ *Use hash trick*